



Hedayah

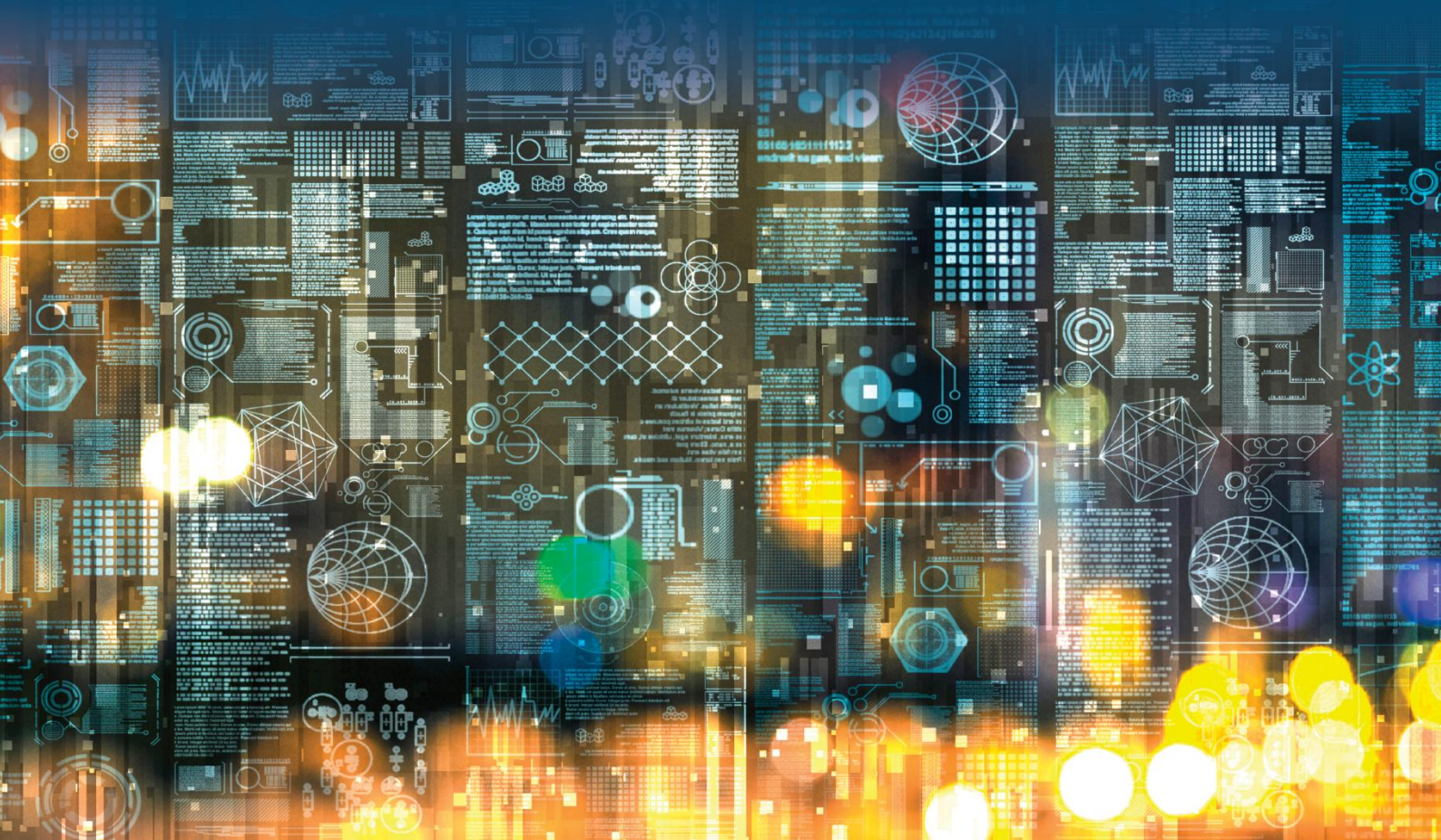
Countering Extremism
& Violent Extremism

RESEARCH BRIEF

Artificial Intelligence for Counter Extremism

Exploring Threats, Challenges,
Opportunities and Needs for Leveraging
Generative AI in Countering Extremism

Anne Craanen, Emma Allen &
Farangiz Atamuradova



The views expressed in this Research Brief are the opinions and work of the authors, and do not necessarily reflect the opinions or views of Hedayah or any of the participating organizations or individuals.

© Hedayah, 2025. All rights reserved.

ABOUT HEDAYAH

Hedayah was created in response to a growing desire from the international community and members of the Global Counter-Terrorism Forum (GCTF) - which now represents 31 countries and the European Union - to establish an independent, multilateral ‘think and do’ tank devoted to countering extremism and violent extremism. Since its inception, Hedayah has evolved into a passionate, driven, and international organization that brings together a vast network of unparalleled experts and practitioners to counter and prevent extremism and violent extremism. Twelve members of the GCTF are representatives of our diverse Steering Board, which provides strategic oversight. As the International Center of Excellence for Countering Extremism and Violent Extremism, we are committed to innovation, neutrality, integrity, diversity, and technical excellence by delivering groundbreaking research, innovative methodologies, and programs. Our approach is to deliver real and sustainable impact to governments, civil society and people impacted by extremism and violent extremism through local ownership and collaboration.

ACKNOWLEDGMENTS

Hedayah expresses its deepest appreciation and gratitude to all who took part in the research, including experts from the following organizations:

- Global Internet Forum to Counter Terrorism (GIFCT)
- University of Silesia
- Paperlab
- Moonshot
- CIVIPOL/Hedayah
- University of Swansea
- Counter-Terrorism Committee Executive Directorate (UN CTED)
- Anglia Ruskin University
- Institute for Strategic Dialogue (ISD)
- Swansea University / VOX-Pol
- Indonesia Knowledge Hub on Countering Terrorism and Violent Extremism
- University of Wisconsin-Madison
- Southeast Asia Regional Centre for Counter Terrorism (SEARCCT)
- Deakin University
- Extremism and Gaming Research Network (EGRN)
- Modulate.ai
- Centinel

Suggested Reference:

Anne Craanen, Emma Allen & Farangiz Atamuradova (2025). *Artificial Intelligence for Counter Extremism: Exploring Threats, Challenges, Opportunities and Needs for Leveraging Generative AI in Countering Extremism*. Hedayah, the International Center of Excellence for Countering Extremism and Violent Extremism, United Arab Emirates (UAE).



TABLE OF CONTENTS

EXECUTIVE SUMMARY	4
GENERATIVE AI AND COUNTERING EXTREMISM.....	6
THREATS, CHALLENGES & OPPORTUNITIES.....	7
LEARNING FROM EXPERTS & PRACTITIONERS: FINDINGS FROM ROUNDTABLE DISCUSSIONS	7
<i>Threats</i>	7
<i>Opportunities</i>	10
<i>Challenges</i>	11
EXPLORING EMERGING RESEARCH: LESSONS FROM LITERATURE REVIEW	13
<i>Threats</i>	13
<i>Opportunities</i>	17
<i>Challenges</i>	19
NEEDS & CONSIDERATIONS: RESPONDING TO AND LEVERAGING GENERATIVE AI.....	21
ADDRESSING RESEARCH AND KNOWLEDGE GAPS	21
CREATING INCLUSIVE SYSTEMS AND AVOIDING MARGINALIZATION	21
PRIORITIZING SAFETY BY DESIGN	22
LEARNING FROM TERRORIST USE OF THE INTERNET (TUI).....	22
LEVERAGING MULTISTAKEHOLDER COLLABORATION.....	23
CREATING ETHICAL AI GOVERNANCE.....	23
BUILDING CAPACITY AND RAISING AWARENESS.....	23
RECOMMENDATIONS FOR RESEARCH, GOVERNANCE & RESPONSE	24
<i>Research & Academia</i>	24
<i>Technology & Industry</i>	24
<i>Policy & Governance</i>	24
<i>Capacity Building & Programming</i>	25
BIBLIOGRAPHY.....	26
ANNEX A. DETAILED METHODOLOGY	30
RESEARCH OBJECTIVES & QUESTIONS	30
RESEARCH APPROACH.....	30
<i>Literature Review</i>	30
<i>Roundtable Discussions</i>	31



EXECUTIVE SUMMARY

This Brief examines what is known about artificial intelligence (AI) and its implications for counterterrorism and counterextremism. The focus is on understanding the threats, opportunities, challenges, and needs associated with AI in the context of terrorism and violent extremism.

The findings and recommendations presented in the article are derived from a targeted review of existing literature and roundtable discussions with experts from academia, technology sectors, government, and civil society. These included five roundtable discussions, organized by Hedayah. For the literature review, a targeted analysis using key terms such as “artificial intelligence” and “terrorism” were used, as well as a more manual search to ensure diversity in region, background, focus, and approach. In total, 52 pieces of literature were examined.

This Research Brief notes key features of the existing research around generative AI and counter extremism, violent extremism and terrorism efforts and applications. **In summary:**

- Terrorist exploitation of generative AI is currently at an experimental level, focusing on multilingual translations, which facilitate quicker dissemination of radical content across different languages and target local grievances for radicalization, content creation, content adaptation, and specifically deepfake creation for disinformation.
- Current research is focused predominantly on predicting, anticipating, and hypothesizing what future use of AI by terrorists and violent extremists will look like rather than current exploitation (or wider considerations about how erosion of trust and confidence in the wider information landscape could affect the factors that drive radicalization and extremism).
- In terms of opportunities to utilize generative AI for prevention efforts, most analysis so far has gone towards using AI for detection and coding extremist content, content moderation such as content removal and distributing counternarratives to audiences assessed as vulnerable based on their online behaviors and evaluating the efficacy of these methods.
- In terms of challenges, literature and experts interviewed note that AI systems often inherit biases present in their training data, which can reflect and perpetuate societal biases such as racism, sexism, and other forms of discrimination. This can lead to skewed outcomes in AI applications, such as facial recognition and predictive policing, disproportionately affecting marginalized communities and reinforcing existing inequalities.

Building on these concerns raised by literature, and the feedback of participating experts, we highlight the following **key takeaways and recommendations** for the way forward in utilizing generative AI for countering extremism:

- **Understanding potential threats or harms:** While generative AI presents positive opportunities, its use by terrorists and violent extremists has potential to exacerbate existing challenges when it comes to preventing and countering extremism. Existing knowledge in relevant areas, from counter extremism to understanding terrorist use of the internet, must inform our response. However, to ensure ethical and effective responses, understanding the potential harms – including gendered harms or harms to marginalized or vulnerable groups – is vital.

- **Ensuring inclusive approaches:** There is a need to promote diversity within AI research and development teams to ensure that AI systems are designed to serve the needs of a diverse global population. This is relevant for both the design and deployment of AI systems.
- **Focusing on safety:** Safety by design is crucial to develop AI systems with transparent decision-making processes and robust mechanisms for accountability. This includes creating clear guidelines for training data and implementing rigorous testing to identify and mitigate biases, ensuring ethical and fair AI deployment that is more likely to serve its potential effectively. This should go together with a human-in-the-loop approach for using AI.
- **Collaborating across disciplines and sectors:** Collaboration between tech companies, governments, and civil society will be pivotal to establish comprehensive safety standards. This includes creating effective reporting mechanisms for harmful content and continuously updating safety protocols in response to emerging threats.
- **Building ‘AI Literacy’ into existing digital, media and information literacy efforts:** Digital and/or media and information literacy can be strengthened by specific efforts on ‘AI literacy’, better enabling individuals to recognize and critically assess AI-generated content. There may be multiple benefits to this - empowering users to differentiate between credible information and mis- or disinformation and thus strengthening their resilience to mis- and disinformation; but also reducing the space for malicious actors to pass as trusted sources, and increasing trust and confidence in individuals engaging with media and information as they feel better equipped to understand the information they encounter. Research is needed to understand how our existing media and information, or digital, literacy efforts can effectively respond and adapt to the emergence of generative AI.
- **Developing practitioner knowledge:** Continuing to build awareness and knowledge of how generative AI functions, the challenges, risks, and opportunities it presents, and how to ethically and effectively utilize this tool for countering and preventing extremism, must be a key focus for counter extremism and violent extremism (CEVE) practitioners moving forward.
- **‘Right-sizing’ our responses:** Ensuring that we ‘right-size’ our response – utilizing existing tools and learnings, and not over, or under, estimating the threats posed by use of generative AI, or what it can and cannot do for prevention practice – is key.
- **Growing the evidence base by examining the development and use of generative AI:** As use of generative AI grows and evolves, there is a need for research to continue in order to move beyond speculation and to further expand evidence-based research that analyses the extent of the problem, what types of online harms are created by artificial intelligence, and how this can be mitigated; and to provide evidence-based approaches for response for practitioners and policymakers.



GENERATIVE AI AND COUNTERING EXTREMISM

By January 2023, artificial intelligence (AI) had emerged as the fastest-growing technology in history, with ChatGPT boasting 180 million monthly active users at the time of writing (McKendrick, 2024). This rapid adoption of generative artificial intelligence (often referred to as ‘generative AI’) has profoundly impacted various sectors, presenting significant opportunities for industries like finance, healthcare, government services and media. However, it has also raised pressing concerns in fields such as security, particularly in counter extremism and counterterrorism.

Initial reactions to the rise of generative AI included alarming predictions about its potential misuse by terrorists and violent extremists. Headlines such as “AI poses national security threat, warns terror watchdog” and “Extremists could use AI to plan attacks, Home Office warns” (Guardian, 2023), and “ChatGPT could promote “AI-enabled violent extremism” (Telegraph, 2023), were some examples of the warnings offered. Whereas generative AI seemed like a novel development, artificial intelligence has been used for decades, both by terrorists and violent extremists, as well as for counterterrorism purposes.

Recognizing the fast-paced nature of technological advancement there is an urgent need for empirical research to supplement speculative warnings. Hedayah, in collaboration with the Global Network on Extremism and Technology (GNET), the research arm of the Global Internet Forum to Counter Terrorism (GIFCT), highlighted this issue in its recent Research Conference events in 2023 and 2024, discussing some of the challenges surrounding generative AI with key experts. Building on this, and on the feedback of its own stakeholders and counterparts, Hedayah saw a need to identify the threats that generative AI poses in terms of terrorist and extremist exploitation, as well as to consider the opportunities generative AI may bring to preventing and countering extremism and terrorism, whilst being aware of the ethical challenges. This is intended to inform Hedayah’s own future programming and research efforts, and to contribute to broader knowledge on the topic.

The research sought to answer the following high-level questions, in line with the Initiative’s objectives:

- **Threats:** What are the key current threats that artificial intelligence, particularly generative AI, poses in terms of terrorist and extremist exploitation, and what new threats may be emerging?
- **Opportunities:** How could generative AI be used, or how is it already being used, to support efforts in countering and preventing extremism and violent extremism? What lessons can we learn from previous uses of (non-generative) AI for these purposes?
- **Challenges:** What are the potential ethical and practical challenges that may be associated with utilizing generative AI for prevention of extremism and violent extremism?
- **Needs:** What needs are emerging or expected in terms of utilizing generative AI for prevention or addressing associated challenges, and how might these needs be addressed?

The research also sought to consider the following cross-cutting issues:

- **Gender differentiated challenges:** How may individual factors, primarily gender, but also age, background, etc. impact these threats, opportunities, challenges, and needs?
- **Non-Western perspectives:** How do current threats, opportunities, challenges and needs differ in different contexts, in particular in non-Western contexts?

To evaluate the key themes related to the exploitation of generative AI by terrorists and the opportunities it presents for counterterrorism, a targeted literature review was conducted, and in total, 52 studies were analyzed. Primary research was also conducted, with five roundtables organized with representatives from the practitioner, academic, tech, civil society, and governmental sectors. Experts were asked to specifically comment on the current threat landscape and exploitation of generative AI by terrorists and violent extremists,



how AI can be used for counterterrorism or counter extremism efforts, what the challenges of doing so could look like, and what the current needs are to counter such exploitation. For more information on the research methodology, please see *ANNEX A. DETAILED METHODOLOGY*.

This Brief aims to synthesize knowledge regarding the threats, opportunities, challenges, and needs associated with generative AI in relation to terrorism and extremism. It will present insights from roundtable discussions with experts from academia, technology, government, and civil society, examining current exploitation patterns, successful content moderation techniques, and future directions. Additionally, it will analyze existing research, highlighting key themes and how factors like identity— for example, gender, race, and background— may shape these dynamics, particularly in non-Western contexts. In doing so, this paper establishes what we know so far and lays out considerations for where we need to go, both in research, but also in programming, governance, and regulation, and concludes with recommendations.

THREATS, CHALLENGES & OPPORTUNITIES

This Brief’s findings represent both the emerging literature on the intersections of generative artificial intelligence and counter extremism and counter terrorism, and expert voices on the issue from practitioners, academics and other key stakeholders, in order to consider the broad spectrum of potential concerns – both positive and negative – that may be relevant to inform practice in efforts to counter and prevent extremism, violent extremism and terrorism.

LEARNING FROM EXPERTS & PRACTITIONERS: FINDINGS FROM ROUNDTABLE DISCUSSIONS

The following section summarizes the main takeaways from the expert stakeholder roundtable discussions, focusing on the threats, opportunities, and challenges they identified.¹ This component of the research was particularly important due to the relatively nascent nature of generative artificial intelligence, and as such, of the research documenting generative AI.

Threats

Taking all observations by the experts from all roundtables, there seemed to be fairly consistent agreement between the types of threats generative AI currently poses. Overwhelmingly, experts agree that terrorist and extremist² use of generative AI is currently at an experimental level, with a lot of predictions and anticipation being indicated by early trends but rarely being grounded in evidence. Experts highlighted that even in contexts where there is limited engagement with generative AI, concern is growing over time. January 2025 brought with it an early example of an ‘AI-assisted’ attack (NBC News, 2025), further highlighting the potential challenge and its relevance. However, it is also important to note that experts also consistently flagged in discussions that much of the discussion taking place around threats from AI is based speculation – some of the threats discussed in qualitative research and in the literature are based on things that AI cannot or has not yet been able to do, or which it can do to some degree but has not yet in these forms. While it is important to consider these threats, as with any risk assessment it is also key to consider their likelihood – and in many cases, there were also reasons why some of threats discussed might not be likely at scale, ranging from the sophistication

¹ This section presents findings from expert roundtable discussion held from August to September. As these were conducted under Chatham House rule, individuals are not attributed.

² Please note that the term ‘terrorist and extremist’ is used in this Brief to refer to actors involved with or supportive of terrorism, extremism, and violent extremism conducive to terrorism.



of current generative AI technology, the capacity of its users, its comparable usefulness with other existing technologies, and the attitudes of terrorist and extremist actors towards these technologies.

In terms of threats, the following themes emerged in the literature:

Propaganda and content can be produced on a larger scale, more quickly and with greater reach

Terrorist and extremist actors focus significantly on propaganda production. Stakeholders expressed concerns about the increasingly high quality and volume of material produced using generative AI, posing challenges for tech companies to detect and moderate such content effectively. The sheer quantity of content could also challenge authenticity and trust, making it harder to differentiate between synthetic and real material, or an increasing cynicism that ‘everything is fake’. A common practice among Terrorist and extremist actors is the use of generative AI for translation, enabling the near-simultaneous creation of content in various languages, which can then be disseminated globally. Experts highlighted targeted hate materials, such as synthetic anti-Muslim or antisemitic content distributed through social media platforms, as well as memes, ironic material, and posters.

Disinformation and misinformation can be enabled by greater speed of production, and by erosion in trust

In the realm of generative AI, distinguishing between terrorist and violent extremist propaganda and disinformation can be challenging. This underscores the increasingly hybridized threat landscape that experts have long warned of, wherein ideologies and organizational affiliations are no longer distinct or mutually exclusive for many actors holding terrorist or extremist views, adding a further layer of complexity to identifying and responding to mis- or disinformation. Disinformation is a key threat posed by generative AI, as it allows terrorist and extremist actors to spread false or inaccurate information on a larger scale, with greater speed and sophistication, creating difficulties for tech companies and regulators to respond. Deepfakes, in particular, were highlighted as a major concern, making it even more difficult for average internet users to distinguish between fact and fiction. Such material and broader disinformation have the potential to further erode trust in official information sources and mainstream media, with significant consequences. It also enables public figures to deny their real statements, attributing them to deep-fake technology. The potential use of AI to influence large datasets, flooding forums or sites used for training by AI with mis- or disinformation that is difficult to identify as terrorist or extremist and thereby mainstreaming extremist narratives, was a concern also raised by experts. Masking or anonymization, including voice masking to make it appear that someone else is speaking, is also highlighted as a worrying tactic for spreading misinformation and disinformation or to discredit individuals or groups. Deepfakes or synthetic media can be used to recreate notable terrorist figures, such as Osama Bin Laden, for recruitment purposes. Voice masking can also allow individuals to seem as though they come from different demographics, furthering dog-whistling.³

Avenues for radicalization and recruitment can be increased through chatbots and translation

The combined effects of the aforementioned factors could have significant consequences for radicalization and recruitment. A notable example frequently mentioned is the Windsor Castle attacker in the United Kingdom, when Jaswant Singh Chail aimed to "kill the queen" on Christmas Day 2021 (BBC News, 2023). Police interviews revealed that part of his radicalization journey was influenced by engaging with a chatbot in which he created a companion named Sarai. He perceived himself to have formed an intimate relationship with the chatbot, and their interactions revealed that the chatbot's agreement with the attacker, as well as its reinforcement of his actions and beliefs, seemed to have contributed to Chail's attempt to harm the Queen of the United Kingdom. Experts argue that chatbots can create a false impression of support, known as “astroturfing”, and while this has applications outside of this context, it may also reinforce radical beliefs, potentially encouraging someone to commit violence.

³ ‘Dog whistling’ is a coded message that is only understood by a particular in-crowd.



The potential for the establishment of highly personalized recruitment through generative AI is also noted, as it allows terrorist and extremist actors to exploit personal grievances and target propaganda according to those grievances as well as language and information preferences. Chatbots are particularly highlighted in this context. While there is consensus on these concerns, experts also agree that more research is needed to understand the exact impact of chatbots on radicalization, as current evidence is limited. They also highlight that, while generative AI content may play a role, the real-world activities of terrorist and extremist actors may still have a greater impact on an individual's radicalization or recruitment process. In addition, they argue that the use cases they highlight are hypothetical and that currently, the development of AI technology is not yet at this level of nuance and sophistication – though it could soon reach this level. It is also possible to draw parallels with recent challenges in countering extremism online, such as the previous impacts of non-generative AI models on radicalization pathways – for example, the role of algorithms, which responded to content preferences of individuals in a way that supported or reinforced extremist or terrorist beliefs by delivering increasingly high volumes of increasing extreme content.

AI can be used for attack planning or operational purposes

Terrorist and extremist actors can also use generative AI for more operational purposes. Experts highlighted the alarming use of generative AI for attack planning and for learning new tactics, such as malicious codes for cyberattacks or the use of drones. Additionally, chemical, biological, radiological, and nuclear (CBRN) materials are available online, and AI can make these more accessible and user-friendly to non-experts. Red-teaming exercises show that extremists are testing AI chatbots, such as using Meta's open-source data to build bots aimed at radicalizing individuals into increasingly extremist beliefs. On a more sophisticated level, manuals on creating AI without ethical and safety guardrails (such as those that block prompts for bomb-making manuals) were also emphasized as an emerging threat.

Overall, experts noted that the cyber threat might be overemphasized, given that most terrorist and extremist actors are not programmers and may not be able to carry out sophisticated attacks - however, they warn that terrorist and extremist organizations may aim to recruit more tech-savvy individuals. Despite the potentially low sophistication, the low barriers to entry mean terrorist and extremist actors can easily adopt new technologies such as generative AI, and as they become more accessible, better trained, or easier to use, this trend may change.

AI poses key risks for trusted actors with limited resources, such as civil society organizations

Beyond the risks posed by terrorist and extremist use of generative AI to global security, experts also highlighted that generative AI could pose a risk to the viability of the civil society organizations that play a vital role in building resilience and countering extremism. The advent and growth of generative AI still has the potential to result in reduced staffing (and thus job losses) and funding changes or shortfalls for CSOs, as it may for other organizations. This would greatly reduce their capacity to serve as key actors contributing to resilience. Further, organizations with more limited resources or capacity may face the highest barriers to being effective users of AI – this may result in a failure to use AI safely and ethically, exposing their data assets to risks of theft or breach. Experts consistently emphasized that AI cannot be used effectively and in a low-risk manner without a good understanding of how it works, and how the specific tool in question uses data, as this may lead to negative unintended consequences both for users and those they work with. Smaller organizations may also be more vulnerable to the malicious use of AI.

Finally, these organizations may be particularly impacted by the general erosion of trust in the information environment that AI stands to create, given that trust and relationships with communities are often the foundation of civil society's work.



Opportunities

In terms of opportunities, stakeholders identified numerous themes. Importantly, they also highlighted that the use of AI for counterterrorism or counter extremism efforts does not come without challenges. References to such challenges will be mentioned in this section as they are important to accurately evaluate the opportunities identified, however the next section will discuss this in more detail.

AI can help to identify radicalization pathways

Experts from the roundtables highlighted that AI can potentially help in detecting individuals' radicalization pathways by identifying key signals along more established trajectories. They specifically note AI's potential to alert patterns that may be indicative of radicalization, and to be used to engage with those individuals in ways that have the potential to serve them with alternative information, content or interactions that may redirect them to more positive behaviors through collaboration with schools, community or home environments. Additionally, AI could be used to test, assess and refine interventions – for example, by having an AI chatbot 'act' like a vulnerable individual and then trialing points of intervention, de-escalation, or disengagement. Though it must be utilized with caution, given the known biases involved, predictive analysis using generative AI may help us to detect when someone is at risk of becoming violent, enabling early intervention. Generative AI can also help with large-scale data analysis to uncover threats and prevent violence, addressing longstanding challenges such as differentiating between so-called "s***posting"⁴ online and genuine threats to violence.

Finally, counter or alternative narratives have been a tactic used for some time, for example, as in Moonshot's *Redirect Method*.⁵ Generative AI may be able to optimize these approaches, especially in resource-constrained contexts, addressing languages and situations previously difficult to manage. Experts highlighted early and emerging cases that suggested AI chatbots should not be the 'sole' responder but could play a role in deradicalization or disengagement efforts – for example, either by acting as a 'safe' source of information for those who did not feel comfortable speaking to another human on this sensitive issue, or by acting as a initial method of engagement and referral. In discussing this future of counter-radicalization bots, experts were optimistic but cautioned that these should be deployed carefully as current AI systems are easy to jailbreak⁶. A compromise could be using chatbots to handle initial queries before referring individuals to human counsellors – however, it is critical to ensure that these human support mechanisms exist and that referral mechanisms are in place for these AI-based approaches to direct to, as otherwise this approach may heighten risk rather than scaling up support.

AI models, including machine learning and NLPs, have already been deployed as content moderators

Artificial intelligence, including machine learning (ML)⁷ and Natural Language Processing (NLP)⁸, has moderated terrorist content on the internet long before the advent of openly-available models such as ChatGPT. The potential of generative AI extends these capabilities. Experts highlight the role of Large Language Models (LLMs) in filtering hate speech and identifying specific forms of it, such as anti-Muslim or antisemitic narratives. However, this currently requires a "human-in-the-loop" – i.e., a human moderator or reviewer – to ensure accuracy and fine-tune these models. However, these models are resource-intensive and have significant financial costs, and as such are used predominantly in contexts where this kind of resourcing is feasible, and have proven to be vulnerable to administrative and commercial decisions by platform owners or regulators. One noteworthy opportunity is that AI may not only be able to optimize content moderation but

⁴ Deliberately posting provocative comments to distract from other conversations.

⁵ For more on Moonshot's *Redirect Method*, see <https://moonshotteam.com/the-redirect-method/>

⁶ Jailbreaking refers to the practice of modifying or removing guardrails in programs or devices created by the designer in order to perform functions that were not intended (or deliberately prohibited) by the designer.

⁷ Machine learning is a form of artificial narrow intelligence (ANI) which equals or exceeds human intelligence for specific tasks (Macdonald et al., 2024).

⁸ NLP are another form of AI based on classification, which includes a large corpus of texts, which have been manually annotated by human reviewers, to train the AI to recognise the features of specified categories to predict whether a new item of content belongs to one of these categories (Macdonald et al., 2024).



simultaneously safeguard human moderators from viewing violent extremist and terrorist content, for example, by providing warnings or blurring graphic material. This would significantly minimize risks to human moderators by protecting them from distressing material and subsequent vicarious trauma or other negative effects they may experience.

AI can do large-scale work quickly, addressing resource limitations

For smaller or less well-resourced organizations or actors, such as civil society actors for example, AI tools becoming more accessible can provide vital support in the face of resource limitations. AI tools can scale up organizations' efforts, potentially allowing for more significant impact. Experts mention Web 4.0 and AI technology experimentation in Africa, showing how AI adoption varies across regions and creates new opportunities.

The preventive potential of Generative AI is still largely unrealized

While stakeholders identify numerous ethical and governance challenges in using AI for prevention, detection, and content moderation, they also note the slow adoption of AI for positive interventions. Despite initial high hopes for generative AI, its implementation has been slow. There are exceptions, such as [Textgain](#) using generative AI to balance datasets for more accurate hate speech detection, but AI's role in counterterrorism has not yet been as widespread as anticipated. Experts reveal that the integration of AI in preventing and countering extremism (P/CVE) programs has been similarly slow as efforts to counter hate-speech.

Challenges

There are numerous challenges that experts have identified, ranging from AI governance, regional biases, gender issues, and ethical concerns.

Biased training data can lead to adverse outcomes

Experts in all roundtables highlighted that there are biases in AI training data, potentially causing multiple practical and ethical problems. For example, AI systems – which have often been trained on data that overrepresents certain racial or gender groups – can lead to inaccurate outputs, such as facial recognition biased towards white subjects, whilst profiling disproportionately impacts people from the Global South, or language generation that overrepresents that tone and manners reflected in the ways that men may speak in public forums, but not the way that women may communicate in more private forums. This issue extends beyond technology, reflecting the assumptions of those who create the technology and datasets. Experts note a lack of diversity in training data, which limits the nuanced understanding of marginalized communities' experiences. Additionally, smaller languages are less understood by AI systems, hindering content moderation and exacerbating the digital divide. Ethical concerns also arise from using certain types of content for training data – for example, while the use of Child Sexual Abusive Material (CSAM) or terrorist content in training data can improve harmful content detection, experts conclude this presents various ethical and legal dilemmas, including by relying on us actively circulating harmful materials.

A critical issue raised across stakeholder roundtables is the lack of transparency in AI model training for content moderation. Transparency can help to create fairer and more equitable training datasets, enabling accountability and oversight. Experts noted that outsourcing content moderation to third-party vendors leads to situations where tech companies do not understand their algorithms' training data or functioning. Transparency can be improved through watermarking⁹ and labelling AI content, but many tech companies have yet to adopt these practices, and they are not yet legally required in many contexts.

⁹ Watermarking is a digital mark that tech companies can add onto data to indicate it was generated by artificial intelligence. Such watermarks can then be picked up by computer systems, creating a record for the source of the content.



Effective oversight requires keeping a ‘human-in-the-loop’

Experts also agree that AI can offer opportunities for the counterterrorism sector, but it must be integrated with a “human-in-the-loop” – meaning, a human actor who reviews or takes certain decisions. AI should complement human decision-making, rather than replacing it. AI systems without this kind of ‘check’ risk causing further harm, with experts noting various past examples of the pitfalls of over-reliance on technology without human input or oversight. For example, in the case of content moderation, experts also warn against replacing human content moderators for cost-cutting purposes, as some platforms have already done (Prada, 2024), and experts highlighted that the risk of false positives (and negatives) necessitates human oversight, as terrorist and extremist actors often adeptly circumvent content moderation, requiring nuanced assessments that generative AI may not ever be able to provide.

Generative AI can both bridge divides and widen them

Wide variations in digital literacy now include “AI literacy”. In regions with low internet penetration, AI adoption often remains limited, further marginalizing already excluded populations. Experts stress the need for critical understanding of AI tools and capabilities, and for integrating AI systems into digital literacy efforts.

Factors of marginalization also impact individual experiences of AI. For example, gender may impact AI use and adoption - women, often victims of harassment, cyberbullying, deepfakes, and gendered disinformation, may be less willing to engage with AI tools, resulting in increased computer and digital literacy gaps. AI literacy also varies across urban, suburban, and rural areas, with rural populations often left out, as seen in countries with uneven internet penetration. Further, “AI literacy” builds on other forms of literacy – media and information literacy, for example, but also basic literacy, as many AI tools have a text-based interface requiring users to be able to read, write, and type to effectively engage with them, and potentially further widening existing gaps in access.

However, it is important to note that generative AI also presents opportunities to bridge divides in access to information – such as linguistic divides, making vital content available in languages previously under-served as translation tools become increasingly sophisticated for a wider range of languages. In this way, AI has the potential to increase information access and subsequently, increase resilience in some ways.

Ethical concerns around privacy and transparency must be taken seriously

Using AI for counterterrorism and counterextremism poses various ethical concerns and the widespread utilization of AI will result in outcomes that could effect the factors that contribute towards grievances and vulnerability to extremism, including perceptions of victimization by the State or the impacts of environmental degradation. Privacy is a significant issue, with experts warning against AI's role in surveillance, potentially leading to mass privacy violations. Further, AI systems require immense computational power, raising environmental impact concerns. Transparency, accountability, and oversight are crucial for ethical AI deployment, ensuring compliance with legal frameworks and human rights.

Regulatory landscapes are still evolving

The legal and policy landscape governing the applications of generative AI as it may relate to terrorist exploitation, and to potential uses for prevention, is still evolving. The recent AI Act¹⁰ in Europe has raised questions about exactly how AI can be used in prevention-related efforts. The Digital Services Act (DSA)¹¹ addresses systemic risks and factors exacerbating them, including algorithmic amplification of illegal content,

¹⁰ Available online at <https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>

¹¹ Available online at https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en



with high-risk AI systems required to detail their workings. Regulators face challenges holding companies accountable, with many firms still non-compliant with legislation.

Experts also expressed concerns about the impacts of such legislation on developing countries and specific sectors like counterterrorism. As regulation often lags behind technology, this can result in ethical and legal grey areas being outsourced to the tech industry, for example, as to what constitutes ‘high risk’ categories and how this applies to companies. However, while regulation is challenging, a lack of AI regulation is also undesirable, as this would leave the industry with no accountability, as many countries still lack legal frameworks for managing data in AI-driven open-source intelligence (OSINT).

Employing ‘safety by design’ in the development and deployment of AI tools is a shared responsibility

Tech experts in the roundtables highlighted that companies often grapple with ethical dilemmas in deploying and monitoring AI tools. While some companies develop internal structures for raising ethical concerns, others do not. "Safety by design"¹² is crucial, ensuring ethical considerations from technology development through to post-production testing. However, not all tech companies secure their technologies, as exemplified by Gab’s *Ayra* model, which denies the Holocaust and lacks safeguards on certain problematic prompts.

EXPLORING EMERGING RESEARCH: LESSONS FROM LITERATURE REVIEW

In addition to the roundtable discussions to hear from experts, the research conducted for this Brief also included a review of existing literature. These have been presented separately to the feedback of experts and practitioners to ensure a clear representation of what is expert opinion and what is captured in literature.

In terms of the overall content and focus of literature, this review found an equal number of articles addressing the threat and exploitation of AI by terrorists and violent extremists and the opportunities of AI for counterterrorism, including topics such as use for counter narratives and content moderation. Many articles commented on ethical and governance issues, as well as biases mostly related to gender and background, but very few focused on this specifically.

Threats

The following section is divided into current threats - arguments made by the literature on existing evidence - and future threats – what authors have predicted or hypothesized that future exploitation of AI could look like.

Current Threats

In keeping with expert feedback in roundtable discussions, the literature also identifies that currently, terrorist exploitation of generative AI is largely experimental (Wells, 2024).

Enabling multilingual content creation and dissemination is a major impact of generative AI

Tech Against Terrorism (TAT) wrote one of the first articles addressing the current exploitation of generative AI, and outlines specific tactics used for propaganda creation (Wells, 2024). As the experts in roundtables concluded, TAT confirmed that terrorist and extremist actors are using generative AI for multilingual translations, to be able to translate material more quickly and release it simultaneously in multiple languages (Tech Against Terrorism, 2023). The literature also highlights that this has consequences for radicalization, given the ability to target communities regarding local grievances in local languages (McDonald, 2023). Other propaganda tactics include media spawning (editing existing propaganda), fully synthetic propaganda (created from scratch using generative AI), variant recycling (reusing old propaganda), and personalized propaganda

¹² ‘Safety by design’ puts user safety and rights at the centre of the design and development of online products and services.



aimed at targeting an individual's specific grievances for recruitment and radicalization purposes. Siegel (2024) provides concrete examples of such propaganda, noting that groups like Daesh are employing generative AI for content creation. However, Siegel also observes that at present, much of the material remains obviously identifiable as synthetic, even when it has not been watermarked. Others have expanded on these findings, noting that real-life attacks and activities carried out by terrorist groups may still exert a stronger influence than synthetic content (Schaer, 2024).

In addition, almost all articles highlight the use of AI to create deep-fakes. A paper by a Working Group from the Global Internet Forum to Counter Terrorism (GIFCT) discusses the threat posed by deepfakes and cites specific instances observed by experts (Engler, 2023). Vergani et al. (2023) emphasize that such deepfakes have the capacity to manipulate public sentiment, posing a significant danger in contexts where the spread of false information can exacerbate tensions. The Australian eSafety Commissioner (2023) adds to this, pointing out that deepfakes often target women, particularly in the form of non-consensual fake pornography involving women in the public eye. Furthermore, the eSafety Commissioner highlights that AI has already been employed to generate synthetic child sexual abuse material (CSAM).

Generative AI has made existing content moderation systems easier to circumvent

Early literature has also warned that by employing generative AI, terrorist and extremist actors are increasingly capable of evading content moderation systems, for example by slightly altering content on a large scale or producing such vast quantities of synthetic content that moderation systems are overwhelmed (Tech Against Terrorism, 2023). Wells (2024) builds on this by arguing that the use of generative AI complicates content hashing—where a digital fingerprint of terrorist content is created and shared through a hash-sharing database, such as the one operated by GIFCT. According to Wells (2024), generative AI can modify the digital hash without significantly altering the actual content, thereby undermining a crucial content moderation strategy.

AI is getting better at interacting with humans

Experts in the roundtables, as well as literature below, all warn for the potential of using chatbots and generative AI systems for radicalization. For instance, an article by Mantello, Ho, and Podoletz (2023) analyzes how chatbots can be tools for radicalization, given that they can create “affect recognition” and exploit the emotional state of users. The authors suggest that individuals may fail to distinguish between digital human-to-human relationships and human-machine interactions, already arguably seen in the case of the Windsor attacker, with transcripts from his interaction with the chatbot indeed showcasing such an ‘affective bonding’. Mantello et al predict that the use of these emphatic and conversational bots will play a larger role in the radicalization of individuals to extremism in future. Interestingly, Vergani et al (2023) also warn of the radicalization effects of bots and highlight their “seductive” capacity, in another parallel with the Windsor case.

This literature review underscores the multifaceted challenges presented by AI technologies in the context of terrorism and violent extremism, including their capacity to evade content moderation, generate and disseminate disinformation, and exacerbate radicalization through tools like chatbots.

Potential Threats

What is very clear from the literature that focusses on the exploitation of AI by terrorist and extremist, or threat actors, is that most articles currently speak about *hypothetical* use cases, somewhat naturally given the relatively recent advent of generative AI technology. While some articles do speak to both, a great deal of attention has been given to predicting or anticipating the *future* threat of terrorist and extremist use of AI. In terms of hypothetical threats, the following themes emerged.

Creating new threats, or supercharging ‘old’ threats?



Most of the studies reviewed agree that one of the most significant risks posed by AI is its potential to optimize, accelerate, or streamline existing threats and attack methods employed by terrorist and extremist actors, both online and offline (UNICRI & UNCCT, 2021). Similarly, Brundage et al. (2018) highlight that AI will make attacks more effective, harder to attribute, and more precise. The National Counterterrorism Centre (NCTC, 2022) in the United States adds that terrorist and extremist actors could use AI to enhance their tactics, techniques, and procedures and to improve the efficiency of their operations. However, AI may also introduce new threats. Brundage et al. (2018) point out that tasks previously impossible for humans could now be accomplished through AI. Additionally, they warn that AI systems used by security sectors could be targeted by malicious actors seeking to exploit their vulnerabilities.

Cybersecurity threats are enhanced by AI

The United Nations Interregional Crime and Justice Research Institute (UNICRI) discusses both the current and future threats posed by AI, specifically outlining "crime-by-design" methods for both cybersecurity and physical attacks (UNICRI, 2020). They identify Denial-of-Service (DoS) and Distributed Denial-of-Service (DDoS) attacks as significant threats, noting that AI can make these attacks more efficient by overwhelming computer systems and rendering them temporarily unavailable. Other cybersecurity threats enhanced by AI include malware, ransomware, password guessing (Engler, 2023), and CAPTCHA-breaking (UNICRI, 2020). These tools can either optimize existing cybersecurity attacks or lead to new types of attacks. Vergani et al. (2023) note that AI can also facilitate fraud and privacy violations, causing serious societal harm. Similarly, Esmailzadeh (2023) warns of the disruption AI can cause to computer networks, potentially resulting in widespread societal consequences.

Physical attacks may be more likely as AI makes terrorist communications harder to intercept and decode

Another risk posed by AI is the ability to better encrypt terrorist communications, making it more difficult for law enforcement agencies to monitor these exchanges. This could have serious consequences for operational aspects such as attack planning (UNICRI, 2020). UNICRI outlines several physical threats, including the use of autonomous vehicles, such as drones, especially those equipped with facial recognition technology, for attacks. UNICRI also warns that AI could be leveraged for various criminal activities by terrorist and extremist actors. These include surveillance, creating fake online identities, morphing passports, and conducting online social engineering to deceive individuals or organizations.

Sandbrink (2023) examines how AI, particularly LLMs and biological design tools (BDTs), could create biosecurity risks if used by malicious actors, arguing that LLMs may lower the barriers to entry for those seeking to develop biological weapons, as AI could assist with tasks that were previously difficult to carry out. Specifically, LLMs could act as virtual lab assistants or autonomous scientific tools, enabling non-experts to carry out laboratory work. This aligns with findings from other fields, as UNICRI and other studies also conclude that AI will enable terrorists and violent extremists to perform tasks that were previously impractical or impossible (UNICRI, 2020). Sandbrink (2023) further warns that BDTs will expand the capabilities of these actors, posing significant risks in the creation of biological weapons.

AI could support terrorist financing in various ways

UNICRI (2020) explores how AI could be used for financing terrorist operations. They suggest that AI could assist in tricking or blackmailing individuals through audio deepfakes, as well as facilitate crypto-trading for financial gain. Experts also flagged that an important trend to monitor was how organized crime groups deployed evolving generative AI tools, as terrorist and violent extremist groups are likely to 'borrow' from these new applications.

AI can be used operationally by Non-State Actors, just as it may be by states

Kreps (2021) examines the potential use of AI by state actors, but also highlights the implications for non-state actors, specifically terrorist groups, noting that AI could be used for optimizing battleground strategies,



predicting outcomes, and spreading disinformation on a large scale, a concern raised in Radicalisation Awareness Network (RAN) meetings (Directorate-General for Migration and Home Affairs, 2023).

Experts warn that the quantity and quality of propaganda generated by AI is likely to increase over time

While the current use of AI in propaganda creation has primarily been at an experimental level, the literature warns of its future potential. Broderick (2023) predicts that AI will lead to an increase in both the quantity and quality of propaganda, with potential new use cases including AI-written terrorist manifestos, AI-powered gamification of extremism, and the rise of deepfakes. He argues that these advances will allow for more targeted recruitment and radicalization of individuals.

Terrorist groups may seek to recruit engineers or technical experts who can build and leverage AI

Experts in stakeholder roundtables debated the current level of sophistication possessed by terrorist and extremist actors in utilizing generative AI for their purposes – similarly, Bazarkina’s (2023) article discusses what it might look like if terrorists and violent extremists begin recruiting tech-savvy individuals or AI engineers into their groups. This would enable these actors to harness the full potential of AI tools.

Overall, it is important to note that whilst these predictions were made by experts and represent credible concerns worthy of consideration, further (and ongoing) empirical review based on evidence is key to ensuring we do not overpredict, or under analyze, the actual use of AI by terrorist and extremist actors and thus fail to accurately understand the nature of the threat and how we can best respond.

The security of current AI systems

Beyond the use of AI systems by terrorist and extremist actors, the literature also discusses the safety of current generative AI systems, particularly chatbots, when exploited for malicious purposes. Several studies provide early empirical explorations of the potential capabilities of terrorist and extremist actors when using chatbots and AI systems, moving beyond speculation and testing their responsiveness. These studies highlight the varying levels of security across different AI platforms, shedding light on the specific vulnerabilities and capabilities that might be exploited by terrorist and extremist actors.

In a pivotal study, Lakomy (2023) examines the use of chatbots by terrorist and extremist actors, comparing different versions of AI models. He concludes that ChatGPT-4 is more resilient to terrorist exploitation than ChatGPT-3.5, and that Bing Chat offers minimal utility to terrorist and extremist actors. Lakomy’s tests found that Bing Chat was particularly adept at triggering security warnings and resisting jailbreaking attempts, which are efforts to bypass content restrictions. Additionally, the study found that chatbots generally outlink to outdated or deactivated domains, with only one instance of a chatbot successfully linking to pro-Daesh content. A concerning capability was the creation of keyword lists that could help locate terrorist content online; however, Lakomy mentions that this is not unique to chatbots and can be done with other forms of technology. Lakomy also warns that while ChatGPT cannot directly quote terrorist content, it can replicate the logic and narratives of such material, posing a risk for the creation of large-scale textual propaganda.

Building on this, Weimann et al. (2024) tested the susceptibility of AI systems to jailbreaking attempts, which allow users to bypass ethical and security guidelines embedded in the models. Similarly to Lakomy (2023), they compared several AI systems, finding that Perplexity had the highest responsiveness (i.e. easiest to jailbreak) to jailbreaking attempts (75%), followed by Nova (54%), ChatGPT-3.5 (53%), ChatGPT-4 (38%), and Bard (31%) (Weimann et al., 2024). Similar to Lakomy’s findings, Weimann also observed that ChatGPT-3.5 was more vulnerable to exploitation than ChatGPT-4. Interestingly, they found that responses from chatbots were often irrelevant or of varying quality and detail, and that hypothetical prompts were more successful in eliciting responses (65%) compared to specific ones (35%). The study also highlighted worrisome use cases, with



chatbots being responsive to operational prompts related to attack planning (30%), tactical learning (61%), and recruitment (21%), suggesting the potential for terrorist and extremist actors to exploit these systems for operational purposes.

Although ChatGPT-4 appears more secure than its predecessor, Yong, Menghini, and Bach identify discrepancies when it comes to language (Yong et al., 2023). ChatGPT-4 performs well against jailbreaking attempts in high-resource languages but falters with low-resource languages. Furthermore, when tested for translating harmful material from low-resource to high-resource languages, ChatGPT-4 handled the translations with ease, highlighting a global safety risk. The authors stress the importance of improving trust and safety efforts across all language platforms to mitigate these vulnerabilities.

He argues that the technical skills required to develop such a tool are beyond the reach of most terrorist and extremist actors. Europol (2023), however, adds that ChatGPT has proven useful in other malicious activities, such as creating phishing attempts, indicating its utility in cybercrime. Wells (2024) further comments on the potential exploitation of chatbots by terrorist and extremist actors, agreeing that while AI systems like ChatGPT could be exploited, as shown above, the concept of a fully functional "terrorist GPT" designed explicitly for radicalization and recruitment seems far-fetched at present.

While many studies focus on the possible uses of AI systems by terrorist and extremist actors, Allchorn (2023) takes a different approach by analyzing how existing far-right extremist groups view generative AI. Through his analysis of communications from groups like Patriotic Alternative, Britain First, Identity England, and Stephen Christopher Yaxley-Lennon (better known as Tommy Robinson) on Telegram, Allchorn finds that these groups largely reject AI, viewing it as a tool that advances a liberal, globalist agenda in opposition to their far-right ideology. Despite this skepticism, the potential of AI might eventually attract these groups, prompting them to adopt the technology for their own purposes, once they see its utility in furthering their goals (Allchorn, 2023). This is further built upon by Baele and Brace (2024) who caution that as AI systems become more publicly available and open-source, the barriers to entry for their use will lower, making it easier for terrorist and extremist actors to adopt AI technologies. This increased accessibility could lead to the more widespread use of AI by malicious actors, posing a significant threat to global security. Experts echoed this, noting that early research and their own monitoring of VE/terrorist spaces online indicates that many of these actors are hesitant to use AI, and may indeed be anti-AI at this time, though this may indeed change as this technology becomes more effective and accessible.

Opportunities

Beyond focusing on the (potential) use cases of terrorist and extremist use of AI, the literature also heavily focusses on using artificial intelligence for content moderation and counterterrorism purposes. It is important to highlight that this is not a new phenomenon. Technology companies have been using AI for content moderation for years, and law enforcement has used AI for surveillance for decades. However, the past five years have seen growing interest in how AI can be used in counterterrorism efforts, particularly generative AI, both online and offline. Several key themes related to AI's role in content moderation have emerged, alongside other potential applications.

AI is a powerful tool to moderate content at large scale and speed

A key focus in AI's use for counterterrorism is its application in content moderation. MacDonald, Wells, and Mattheis (2024) identified two main techniques: matching-based systems, which use hashing to compare new images against a database of known terrorist content, and Natural Language Processing (NLP), which uses AI and machine learning to classify text. Both systems are more effective when combined with human oversight (i.e., keeping a 'human in the loop', as highlighted in earlier sections of this Brief). Naseer and Shaheen (2023) further explore NLPs use in identifying and countering terrorist content online, emphasizing how advanced algorithms can even predict terrorist actions, allowing law enforcement to intervene pre-emptively. Schroter



(2020) adds that NLPs are crucial for translating minority languages and supporting moderation on niche platforms, while manipulated search engines and recommendation systems can promote moderate content.

Barnes (2022) offers a different perspective, arguing that tech companies' investment in AI for content moderation is driven by their broader goal of scaling up social media platforms. The study suggests that AI allows platforms to maintain growth while automating content moderation, positioning AI as both a solution and a challenge. Fernandez and Alani (2021) highlight two key limitations in AI-driven content moderation: the absence of control groups, which makes it difficult to gauge the impact on radicalization, and issues with data collection and verification. Heller's (2020) evaluation of GIFCT's hash-sharing database (HSB) finds it effective but also subject to technical limitations and raises concerns about free speech.

A related theme in content moderation is the creation of counternarratives. Tekiroglu et al. (2022) examine AI systems that generate counternarratives, finding that autoregressive models like GPT-2 are adept at producing novel counternarratives, while deterministic models are better suited to generating general ones. Montasari (2024) also notes AI's role in detecting extremist content, while Costello et al. (2024) demonstrate how AI chatbots can help counter conspiracy theories, showing a significant decrease in belief among 20% of users exposed to AI-generated counternarratives in their study.

Gandhi raises a crucial issue about the content generated by terrorist and extremist actors using generative AI: much of it falls into a category of borderline content that may not be explicitly terrorist but still causes harm (Gandhi, 2024), and advocates for the development of a taxonomy of harms to ensure the right moderation approaches are applied to this content.

Finally, Macdonald et al (2024) also emphasize the role of AI in safeguarding. AI tools can protect content moderators, researchers, civil society members, and practitioners from the harmful effects of graphic or disturbing content.

Counterterrorism concerns and applications are more extensively researched

While counterterrorism concerns and applications were not a primary focus of this research, as they inform potential threats, opportunities and frame the needs for prevention and response, they were included in this review.

The literature heavily focusses on AI's potential to predict terrorist attacks. Huamani et al. (2020) find that machine learning techniques were able to visualize and predict terrorist attacks, and with the help of classification models, the authors demonstrated that it is possible to predict the type of attack and in which region this may occur. Khan et al (2023) also explore classifiers for predicting the type of weapon used in attacks, achieving high accuracy through various models. McKendrick (2019) adds that whereas AI may have a bad reputation for harming human rights, AI may allow for more effective counterterrorism whilst also ensuring transparency, proportionality and human rights are respected, emphasizing that current regulations may hinder AI's effectiveness without necessarily improving rights protections.

Another theme in the literature is to use AI models for identification of individuals. The article by Tosin et al. (2022) analyses how algorithms can be used for border security, by using facial recognition and comparing them to existing databases. Little consideration however went to the ethical implications of such software and surveillance. Stokes (2021) adds to the use of AI for facial recognition but highlights that AI ethics should be considered before AI is used for counterterrorism purposes.

AI can also counter cyberterrorism. Jha et al. (2023) find that AI can minimize cyberattacks in real-time, enhancing the response of security agencies; however, they also caution that AI-driven cybersecurity measures come with legal challenges, requiring compliance with relevant legislation.



AI's potential for military application is another area of focus. Ceballos (2022) finds that AI systems can improve and complement decision-making processes, and if used right, could protect civilians and reduce casualties, through more accurate targeting.

Countering of terrorist financing using AI is still under-researched

A seemingly understudied trend in the literature analyzed is that of using AI tools to counter terrorist financing. Deloitte's analysis highlights how AI tools can assess high-risk jurisdictions, identify suspicious financial activities, and refine screening processes for politically exposed persons (PEPs) and sanctioned organizations. Machine learning models can be trained to recognize suspicious behaviors and improve name screening accuracy by learning from existing systems' indicators. Of note, is that this is an article written by a corporate organization and within academic literature, this remains an understudied opportunity.

AI offers opportunities to advance the Women, Peace & Security (WPS) Agenda

UN Women and the UN University Institute in Macau (2024) highlight that AI offers opportunities to advance the Women, Peace, and Security (WPS) agenda, noting how AI tools such as social media, chatbots, and mobile applications could usefully raise awareness of WPS issues. Additionally, AI can help foster gender-responsive peace efforts in line with WPS commitments, though further research is needed to explore these possibilities. The literature shows how AI plays an increasingly pivotal role in counterterrorism, from content moderation and safeguarding to predicting terrorist actions and countering terrorist financing. However, ethical concerns, technical limitations, and legal challenges must be addressed to ensure AI's effective and responsible use. The following section will go towards explaining these challenges in greater detail.

Challenges

The literature highlights multiple challenges related to AI's impact on peace and security, particularly in terms of gender-responsiveness, the spread of disinformation, biases, and governance gaps. Addressing these challenges requires inclusive governance, ethical AI development, and stronger global coordination.

AI can create and exacerbate biases

Studies identified several issues when it comes to gender biases and the use of AI tools for both nefarious and positive purposes. First, women face challenges in gaining equal opportunities to lead and meaningfully participate in the design of AI systems (UN Women and the UN University Institute in Macau, 2024). This lack of representation can lead to AI systems that are not inclusive or gender-responsive, impacting the Women, Peace, and Security (WPS) agenda, as well as the AI systems being inherently biased. Second, AI systems, particularly those used in social media, chatbots, and mobile applications, can perpetuate or amplify harmful gendered narratives, such as misogyny and other forms of discrimination. These biases pose significant risks to advancing gender equality in peace efforts. This was also highlighted by the Australian eSafety Commissioner (2023) in regard to pornographic deepfakes of women in public positions. Third, many AI systems are developed and deployed without assessing their gender or human rights impacts, making them potentially detrimental to peace and security, especially in conflict-sensitive environments.

Beyond gender biases, many AI models are primarily trained on Western datasets, potentially leading to biases that may make them less effective, or even harmful, in non-Western contexts. These biases could hinder the success of AI-assisted peacebuilding and counter-extremism efforts in diverse regions (Vergani et al., 2023). This aligns with the findings of Yong et al (2023), who concluded that ChatGPT-4 was trained on high-resource languages rather than low-resource ones, creating risks for all users. Royer (2020) adds to this and argues that too much focus on merely the AI models ignores the social, political, and cultural forces at play in the design, use, and implementation of these systems. Conscious or unconscious level, the academic, economic, cultural,



and social patterns that exist in the contexts where these tools are designed and developed may in fact reinforce existing global inequalities and economic disparities.

Finally, when it comes to utilizing AI for counterterrorism, Stokes (2021) adds that the potential negative or harmful consequences of using AI for counterterrorism can include embedding prejudicial biases that affect minority communities with far-reaching consequences. When AI systems are trained on biased data, this can lead to discriminatory outcomes – for example, over removal and censorship from automated moderation systems, which disproportionately impacts individuals from the Global South (Macdonald et al., 2024).

Governance challenges remain in the face of rapidly evolving technology and use

AI is evolving rapidly, but governance frameworks are often lagging. Ensuring that AI is governed inclusively, for the public interest, and in alignment with international human rights laws remains a significant challenge (High-level Advisory Body on Artificial Intelligence, 2024). While efforts are being made to harmonize standards and create risk management frameworks, there are significant obstacles to achieving global interoperability and coordination of AI governance, which is needed to manage AI's societal risks effectively (Royer, 2020). In addition, governance should include women, engineers with diverse backgrounds, and human rights experts, as well as civil society, to ensure it is ethical and inclusive (UNICRI, 2020).

Accountability mechanisms are limited

The literature finds, and experts agreed, that social media platforms where AI plays a significant role in content moderation and curation often lack adequate accountability mechanisms. This can lead to the unchecked spread of harmful content without sufficient recourse for users (UN Women, 2024; Heller, 2020). The lack of user agency in choosing AI-driven platforms or providers can leave individuals exposed to the risks of bias, misinformation, and manipulation, without the ability to easily opt for safer or more rights-based alternatives. While more open-source AI tools may provide for more diversity and user agency to choose a certain platform (Heller, 2020), this may on the other hand also risk the creation of non-ethical AI systems that lead to further exploitation of these services, as mentioned by experts. However, it is important to note that there is current regulation that can be used to respond to harmful content and create accountability.

Civil society may lack the resources and training to leverage generative AI

Vergani et al (2023) find that CSOs often lack the advanced training and resources needed to use AI responsibly and effectively, and that without support from data scientists, CSOs may struggle to harness AI for counter-extremism and peacebuilding efforts. This mimics some of the observations shared in the roundtable sessions, where the adaptation of AI by CSOs for PCVE measures has been less widespread than initially anticipated. While generative AI can provide creative tools for counter-narratives and digital literacy, using these tools without clear ethical standards and transparency can lead to further harm, including the reinforcement of biases or unintentional misuse (High-level Advisory Body on Artificial Intelligence, 2024). AI systems are prone to evolving challenges, such as biases in algorithms and vulnerabilities in security, which can be exploited by malicious actors (Barnes, 2022). Ensuring responsible AI use requires continuous attention to these risks.

AI has the potential to undermine trust and impact social cohesion

The misuse of AI in spreading disinformation and harmful narratives can undermine trust in institutions and erode social cohesion, making peacebuilding efforts more difficult (Gandhi, 2024). Not everyone trusts AI technology (Allchorn, 2023), and its use by governments may further alienate the public from public institutions (Burton, 2023). Burton (2023) finds that when discussing the securitization of AI, AI's involvement in counterterrorism has led to increased political polarization, particularly as AI is framed as a tool of security, and the use of AI in such a way can contribute to the creation of extreme viewpoints and divisions within society, which exacerbates political violence and radicalization - beyond merely examining these malicious use cases of the technology, the social negative impacts from the technology itself warrants more attention.



NEEDS & CONSIDERATIONS: RESPONDING TO AND LEVERAGING GENERATIVE AI

This Brief stresses the necessity of a fair, open, and ethical strategy for using AI to combat extremism, violent extremism and terrorism. By actively considering these factors, governments, the tech sector, and civil society can use AI to its full potential while preserving privacy rights, individual liberties, and moral standards. Ethical AI use is essential to creating a more secure and safe online (and offline) environment for all (Naseer & Shaheen, 2023).

Reflecting on the literature discussed and the inputs of experts, this section of the Brief now turns to a reflection of what we know so far, and where we may need to go.

Addressing Research and Knowledge Gaps

There is a pressing need for **further empirical research to address the gaps in understanding how AI is misused by terrorists and violent extremists**, grounded in evidence. This includes moving beyond speculative studies to quantify the scale of misuse and identify how AI tools are applied across different ideologies (Allchorn, 2023). When conducting such research, it is vital that ethical considerations are kept in mind. Conway has long argued that researchers have a duty to report any terrorist content they find on platforms, a principle that should similarly apply to generative AI content (Conway, 2021). When attempting to jailbreak AI platforms, researchers must exercise caution when detailing the prompts used, ensuring that they do not inadvertently teach malicious actors how to exploit artificial intelligence.

Noting what this Brief has identified in terms of the potential threats which generative AI may pose for marginalized groups, exacerbating biases and creating greater access gaps, further research in particular is warranted regarding its gendered impacts. The literature review identified limited articles referring specifically to the AI-driven harms faced by women, such as pornographic deepfakes (which as of 2019, made up the overwhelming majority of deepfakes (Reissman, 2023)) and non-consensual sharing of imagery, highlighting a largely gender-specific harm resulting from AI that warrants attention through further research. This issue, and other gendered harms, might be addressed by creating a taxonomy of online harms that includes technology-facilitated gender-based violence (TFGBV¹³).

Specific future research needed identified by literature and experts also included further studies on the radicalization effects of chatbots and generative AI (Lakomy, 2023; Weimann et al., 2024), and while some of the studies highlighted in this article have made a noteworthy start, additional research should explore the affective and emotional effects of generative AI systems and how these may impact radicalization and deradicalization. Another area for study maybe the examination of how terrorist and extremist actors could use AI to create not only terrorist propoganda but also extremist content, hate speech, borderline content, and other forms of harmful online material (Gandhi, 2024).

Creating Inclusive Systems and Avoiding Marginalization

As well established in this Brief, AI systems are often trained on Western culture, people, and assumptions.¹⁴ This not only creates an insecure AI environment—given terrorist and extremist actors' ability to exploit lesser-known languages and local grievances and dynamics—but also exacerbates global inequalities.¹⁵ Wealthy nations must find ways to support non-Western countries or states with lower levels of internet penetration and digital literacy to prevent technological advancements from widening the digital divide (Hedayah, 2024).

¹³ For more on TFGBV, see <https://www.unfpa.org/TFGBV>

¹⁴ See for example: Burton, 2023; UNICRI, 2020; McDonald, 2023; Wells, 2024; Macdonald et al., 2024; Hedayah (2024).

¹⁵ See for example: Yong et al., 2023; UNICRI, 2020; Vergani et al., 2023.



Additionally, when AI is used for counterterrorism—whether through content moderation or facial recognition—it may disproportionately impact marginalized communities (Stokes, 2021), securitizing and further marginalizing them. It is vital that actors seeking to use generative AI tools for such purposes have considered and sought to mitigate these risks.

Furthermore, AI systems may alienate some users – for example, women may feel unsafe using AI systems that have been employed to target them (eSafety Commissioner, 2023), and non-Western nations and communities often have less opportunity to engage with certain AI platforms (Burton, 2023). However, discussions on broader gender and identity factors are notably absent - it is crucial to consider to what extent AI systems may propagate harmful narratives, and who determines the assumptions that shape this technology, in order to ensure that AI platforms are open and welcoming spaces for all.

Prioritizing Safety by Design

Experts and literature consistently highlight the importance of a human-in-the-loop approach when using AI systems (Wells, 2024; Macdonald et al., 2024) - this is essential to ensure that classifiers are trained and refined to identify terrorist content across global languages, effectively keeping pace with content moderation avoidance techniques, humor, and nuance, as well as to ensure that AI-driven content moderation does not adversely affect marginalized communities. As some have already suggested, the question should not be how AI can replace humans, but rather how AI can assist humans.

Moreover, Safety by Design is one of the most frequently mentioned recommendations throughout the literature, as well as being identified by the experts interviewed. As articulated by the eSafety Commissioner (2023) in Australia, safety should be prioritized at all stages, from the business case and data selection to model training, refinement, release, user engagement, AI generation, and feedback. While safety-by-design has been mentioned throughout, it is crucial that future research gives more consideration to specific measures and methods for its implementation. The security sector should engage more frequently with computer science departments, particularly concerning this topic.

Although not explicitly mentioned in the literature, reporting mechanisms on major generative AI platforms are often difficult to locate. This complicates the ability of researchers, practitioners, and users to flag illegal or harmful content to these platforms. Given the lessons learned from the exploitation of social media by terrorist and extremist actors, clear and user-friendly reporting mechanisms are critical in countering terrorist and extremist use of social media; this necessity is often codified in legislation such as the Digital Services Act (2022)¹⁶ and the Online Safety Act (2023)¹⁷. AI companies should learn from these examples and establish accessible reporting mechanisms tailored to the specific functionalities of their platforms and the potential nefarious uses to which they can be put.

Learning from Terrorist Use of the Internet (TUI)

While generative AI presents a new and unique tool, and new challenges and opportunities as a result, new technologies are not unprecedented, and the counter terrorism and counter extremism sectors, alongside tech and government, have learned many lessons that are potentially applicable today. It is critical that we learn from the lessons derived from the past decade of exploitation of the internet by violent extremists and terrorists (Wells, 2024). Rather than perceiving AI or generative AI as an entirely different environment capable of immediate radicalization and recruitment, we should be cautious in isolating this threat as unrelated to others. Moreover, as with other technological developments, AI is not a “magic” solution that will eliminate terrorist or otherwise harmful content from the internet. Indeed, the use of AI in content moderation and counterterrorism may share the unintended consequences already observed from content moderation practices, necessitating a holistic approach to address the risks and opportunities presented by AI. While we

¹⁶ Available online at <https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package>.

¹⁷ Available online at <https://www.legislation.gov.uk/ukpga/2023/50>.



must take care to question our assumptions as they apply to this new area, it is also key to consider the various regulatory, legal, policy, and programming tools we already have to respond, and how we can use them.

Leveraging Multistakeholder Collaboration

Both the literature and interviews with stakeholders highlight the need for multistakeholder collaboration. Civil society organizations should receive support from other sectors, such as technology, to fully harness the potential that AI has for their work. Furthermore, tech companies should incorporate academic and civil society perspectives into their safety-by-design processes when building and refining technology, thereby mitigating unforeseen consequences (Hedayah, 2024). Finally, collaboration between governments and tech companies is essential. Although these two entities are often perceived as being at opposing ends, AI technology is here to stay. While regulation is a positive step toward holding tech companies accountable, governments should also learn from tech companies regarding the sophistication and functionality of AI tools. As experts suggest, regulation often lags behind technology, and the only way for regulation and AI development to progress in tandem will be through collaboration.

Creating Ethical AI Governance

AI governance concerns the whole of society, and the involvement of a diverse range of stakeholders is essential to ensure it works for all. AI governance for humanity is therefore not something that should be decided by Western actors - the governance of AI is a global concern, and it should be governed inclusively (High-level Advisory Body on Artificial Intelligence, 2024). Furthermore, there is an urgent need for clarity on the ethical and legal boundaries surrounding AI applications, ensuring that beneficial uses are promoted while preventing potential misuse.¹⁸ As AI technologies rapidly evolve, regulations must also adapt to address emerging threats associated with terrorism and extremism, steering clear of vague guidelines that may lead to confusion and inadequate responses.

Moreover, governments should avoid the pitfalls of overregulation, which can misuse AI as a pretext for excessive control and oversight, ultimately hampering innovation. Establishing international standards is crucial for fostering equitable governance (High-level Advisory Body on Artificial Intelligence, 2024) (High-level Advisory Body on Artificial Intelligence, 2024), with wealthier nations playing a vital role in assisting less-resourced countries to build their capacities. Improved communication between governments and technology companies is equally important, as it encourages a constructive feedback loop that enhances mutual understanding and leads to more effective regulatory frameworks without stifling innovation.

Building Capacity and Raising Awareness

Most articles and experts highlight the need for digital literacy and AI literacy to combat the effects of significant volumes of disinformation, enabling individuals to distinguish between synthetic and credible sources.¹⁹

There is a pressing need to adapt current digital literacy and media and information literacy campaigns or initiatives for the new global context – such approaches should be reviewed, and if necessary updated, to ensure that they reflect the realities of generative AI and the challenges it creates. We do not yet know how well suited our existing toolkits for combatting misinformation and disinformation, and AI will require us to not reinvent, but at least refine, such approaches.

Additionally, as both the capability of generative AI evolves, and the understanding and use of AI develops alongside it both by malicious actors and those working in prevention and response, understanding the specific applications of individual tools, the risks associated with using them, and the benefits of doing so effectively will be vital leveraging AI and responding to the threats and opportunities it presents.

¹⁸ See for example: High-level Advisory Body on Artificial Intelligence, 2024; Royer, 2020; UNICRI, 2020; Heller, 2020.

¹⁹ See for example: Vergani et al, 2023; Engler, 2023; Gandhi, 2024.



RECOMMENDATIONS FOR RESEARCH, GOVERNANCE & RESPONSE

This section concludes the Brief by providing clear, high-level recommendations for various actors who must be part of coordinated, ethical and effective responses to AI from a counter extremism perspective. These include researchers, practitioners, the tech sector, policymakers, and programming actors. While these recommendations are not exhaustive, they highlight key principles for engagement in this space.

Research & Academia

- **Conduct Empirical Research:** Prioritize empirical studies to understand the misuse of AI by terrorists and violent extremists, focusing on quantifying the scale of misuse and identifying the application of AI tools across different ideologies.
- **Understand Gendered Harms:** Expand research on gendered misuse of AI, particularly concerning technology-facilitated gender-based violence (TFGBV) and establish a taxonomy of harms related to online violence, including gendered deepfakes and disinformation.
- **Examine AI's Impact on Radicalization:** Further investigate the potential radicalization effects of chatbots and generative AI, including emotional impacts and the nature of extremist content created by AI tools.
- **Develop Ethical Reporting Practices:** Encourage researchers to report terrorist and extremist content found during AI studies, ensuring ethical considerations are upheld throughout the research process.

Technology & Industry

- **Human-in-the-Loop Systems:** Implement human-in-the-loop methodologies in AI systems to refine classifiers for identifying terrorist content, ensuring nuanced understanding across diverse languages and cultures.
- **Safety by Design:** Adopt safety-by-design principles throughout all phases of AI development, incorporating insights from security experts and computer science departments to mitigate risks effectively.
- **Accessible Reporting Mechanisms:** Develop clear and user-friendly reporting mechanisms for AI platforms, facilitating the easy flagging of harmful content by researchers, practitioners, and users.
- **Invest in collaboration:** Support civil society organizations where possible by upskilling them and sharing best practice for using AI technologies.
- **Transparency:** Provide more transparency on training data, how biases are mitigated, the amount of synthetic material, and other key factors that may support the wider sector to use AI in for positive opportunities and mitigate harmful misuse of the technology.
- **Inclusive Voices in AI Development:** Include diverse voices—women, LGBTQIA+ individuals, and communities from various backgrounds—in the design and development of AI systems to ensure inclusive and equitable platforms.

Policy & Governance

- **Establish Clear Definitions and Guidance:** Create clear, adaptable guidance for current AI regulation with an emphasis on defining ethical and legal boundaries, while avoiding overregulation that stifles innovation. While legislation does exist, it is not uniformly implemented or applied.



- **International Collaboration:** Foster international standards for equitable governance, ensuring well-resourced nations, whether in terms of relevant expertise or funding, support less-resourced countries in building capacities related to AI and its governance.
- **Multistakeholder Partnerships:** Encourage collaboration among civil society organizations, tech companies, and governments to leverage the full potential of AI while mitigating unforeseen consequences.
- **Invest in Both Human Capability and AI:** Continue to invest in human capability as well as generative artificial intelligence to ensure that generative AI is supported where needed by human services that can build on AI-based processes and monitor its outcomes; and recognize where human services will continue to be necessary or more effective, ensuring capacity to provide these services are maintained (for example, ensuring that civil society organizations continue to be capacitated to provide support services to individuals initially identified or ‘referred’ by AI chatbot services).

Capacity Building & Programming

- **Build on Lessons from Previous Responses:** Learn from the exploitation of the internet by terrorists and violent extremists to inform the approach towards AI, recognizing that it is not a panacea for harmful content, to ensure responses are calibrated to the challenge.
- **Monitor and Adapt:** Establish mechanisms for continuous monitoring and assessment of AI tools used in counterterrorism or counter extremism to ensure they adapt to emerging threats and remain effective.
- **Enhance Digital Literacy:** Promote digital and AI literacy programs to equip individuals with the skills to discern credible sources from disinformation, tailored to the challenges presented by generative AI for all age groups.
- **Strengthen Knowledge and Provide Resources:** Work with practitioners, civil society and academia to provide resources and training on how AI can be utilized in prevention efforts ethically and effectively.



BIBLIOGRAPHY

In order of appearance:

- McKendrick, Joe. (2024) "AI 'Fastest-Growing Technology We've Seen In The History Of Our Company'", *Forbes*, 19 September. <https://www.forbes.com/sites/joemckendrick/2024/09/19/ai-fastest-growing-technology-weve-seen-in-the-history-of-our-company/#:~:text=The%20consultancy's%20survey%20of%20%2C800,consulting%20at%20Accenture%2C%20told%20me>.
- Townsend, Mark. (2023) "AI poses national security threat, warns terror watchdog." *The Guardian*, 4 January, <https://www.theguardian.com/technology/2023/jun/04/ai-poses-national-security-threat-warns-terror-watchdog>
- Syaal, Rajev. (2023). "Extremists might use AI to plan attacks – home office warns." *The Guardian*, 18 July. <https://www.theguardian.com/politics/2023/jul/18/extremists-might-use-ai-to-plan-attacks-home-office-warns>
- Hymas, Charles (2023). "ChatGPT could promote 'AI-enabled' violent extremism." *The Telegraph*, 9 April. <https://www.telegraph.co.uk/news/2023/04/09/chatgpt-artificial-intelligence-terrorism-terror-attack/>
- Jones, Ja'han (2025). New Year's incidents highlight the rise of AI-enabled terror. NBC News. <https://www.msnbc.com/the-reidout/reidout-blog/cybertruck-las-vegas-new-orleans-artificial-intelligence-terror-ai-rcna186857>
- Singleton, Tom. Gerken, Tom. & McMahon, Liv. (2023). "How a chatbot encouraged a man who wanted to kill the Queen". *BBC News*, 6 October. <https://www.bbc.co.uk/news/technology-67012224>
- Prada, Luis. (2024). "TikTok Laying Off Hundreds of Content Moderators, Replacing Them With AI," *Vice News*, 14 October. <https://www.vice.com/en/article/tiktok-content-moderation-layoffs-ai/>
- Wells, David. (2024). "The next paradigm-shattering threat? Right-sizing the potential impacts of generative AI on Terrorism". *Middle East Institute*, March 2024. <https://mei.edu/sites/default/files/2024-03/Wells%20-%20The%20Next%20Paradigm-Shattering%20Threat%20RightSizing%20the%20Potential%20Impacts%20of%20Generative%20AI%20on%20Terrorism.pdf>
- Tech Against Terrorism. (2023) Early Terrorist Adaptation of Generative AI. <https://techagainstterrorism.org/news/early-terrorist-adoption-of-generative-ai>.
- McDonald, Broderick. (2023). "The Use of AI in Disinformation & Extremism: Separating Fact from Fiction", December 11. <https://www.oxdisinformationextremismmlab.com/research/article/the-use-of-ai-in-disinformation-extremism-separating-fact-from-fiction>
- Siegel, Daniel. (2024). "AI Jihad: Deciphering Hamas, Al-Qaeda and Islamic State's Generative AI Digital Arsenal", *Global Network on Extremism and Technology*, 19 February. <https://gnet-research.org/2024/02/19/ai-jihad-deciphering-hamas-al-qaeda-and-islamic-states-generative-ai-digital-arsenal/>
- Schaefer, Cathrin. (2024). "How extremist groups like 'Islamic State' are using AI." *DW*, 7 October. <https://www.dw.com/en/how-extremist-groups-like-islamic-state-are-using-ai/a-69609398>
- Engler, Maggie. (2023). "Considerations of the impacts of generative AI on online terrorism and extremism". *GIFCT*, September 20. <https://gifct.org/wp-content/uploads/2023/09/GIFCT-23WG-0823-GenerativeAI-1.1.pdf>
- Vergani, Matteo. Lukman, Barton, Nadia Greg & Zaman, Dina. (2023). "Generative artificial intelligence and countering violent and hateful extremism", *Southeast Asian Network of Civil Society Organisations*. <http://www.sean-cso.org/wp-content/uploads/2023/10/Report-SEAN-CSO-Generative-AI-CVE.pdf>
- eSafety Commissioner. "Tech Trends Position Paper", *Australian Government*. <https://www.esafety.gov.au/sites/default/files/2023-08/Generative%20AI%20-%20Position%20Statement%20-%20August%202023%20.pdf?v=1725311249678>



- Mantello, Peter, Tung Manh Ho, and Lena Podoletz. (2023). "Automating extremism: Mapping the affective roles of artificial agents in online radicalization." *The Palgrave handbook of malicious use of AI and psychological security*, pp. 81-103. Cham: Springer International Publishing.
- Interpol. (2024) "Grooming, radicalization and cyber-attacks: INTERPOL warns of 'Metacrime'" *Interpol*, January 18.. <https://www.interpol.int/en/News-and-Events/News/2024/Grooming-radicalization-and-cyber-attacks-INTERPOL-warns-of-Metacrime>
- UNICRI. (2020). "Algorithms and Terrorism: The Malicious Use of Artificial Intelligence for Terrorist Purposes". *UNICRI, UNCCT*. <https://unicri.it/News/Algorithms-Terrorism-Malicious-Use-Artificial-Intelligence-Terrorist-Purposes>
- Brundage, Miles, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe et al. (2018). "The malicious use of artificial intelligence: Forecasting, prevention, and mitigation." *arXiv preprint arXiv:1802.07228*.
- NCTC. (2022). "Emerging Technologies May Heighten Terrorist Threats." *NCTC*, 14 October. https://www.odni.gov/files/NCTC/documents/jcat/firstresponderstoolbox/134s_-_First_Responders_Toolbox_-_Emerging_Technologies_May_Heighten_Terrorist_Threats.pdf
- Esmailzadeh, Yaser. (2023) "Potential Risks of ChatGPT: Implications for Counterterrorism and International Security." *International Journal of Multicultural and Multireligious Understanding (IJMMU) Vol 10*.
- Kreps, Sarah. (2021). "Democratizing harm: Artificial intelligence in the hands of nonstate actors". *Brookings*, November. <https://www.brookings.edu/articles/democratizing-harm-artificial-intelligence-in-the-hands-of-non-state-actors/>
- Directorate-General for Migration and Home Affairs (2023). "RAN C&N 'What's going on online? Dealing with potential use of deepfakes by extremists', Helsinki 10-11 November 2022". European Commission, 6 March.
- Bazarkina, Darya. (2023). "Current and future threats of the malicious use of artificial intelligence by terrorists: psychological aspects." In *The Palgrave handbook of malicious use of AI and psychological security*, pp. 251-272. Cham: Springer International Publishing.
- Sandbrink, Jonas B. (2023) "Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools." *arXiv preprint arXiv:2306.13952*
- Lakomy, Miron. (2023) "Artificial intelligence as a terrorism enabler? Understanding the potential impact of chatbots and image generators on online terrorist activities." *Studies in Conflict & Terrorism*: 1-21.
- Weimann, Gabriel. Pack, Alexander, Rapaport, Gal, Scheinin, Joelle, Diaz, David and Sulciner, Rachel. (2024). "Generating Terror: The Risks of Generative AI Exploitation". *International Centre for Counter-Terrorism (ICT)*, January 18.
- Yong, Zheng-Xin, Cristina Menghini, and Stephen H. Bach. (2023). "Low-resource languages jailbreak gpt-4." *arXiv preprint arXiv:2310.02446*.
- Europol. (2023). "ChatGPT - The impact of Large Language Models on Law Enforcement." *Europol*, 27 March. <https://www.europol.europa.eu/cms/sites/default/files/documents/Tech%20Watch%20Flash%20-%20The%20Impact%20of%20Large%20Language%20Models%20on%20Law%20Enforcement.pdf>
- Allchorn, William. (2023). "Global Far-Right Extremist Exploitation of Artificial Intelligence and Alt-Tech: The Cases of the UK, US, Australia and New Zealand." *RSIS*, <https://www.rsis.edu.sg/ctta-newsarticle/global-far-right-extremist-exploitation-of-artificial-intelligence-and-alt-tech-the-cases-of-the-uk-us-australia-and-new-zealand/>
- Baele, Stephane J and Brace, Lewys. (2024). AI Extremism: Technologies, Tactics, and Actors. *Vox-Pol*, <https://voxpoleu/wp-content/uploads/2024/04/DCUPNO254-Vox-Pol-AI-Extremism-WEB-240424.pdf>
- Macdonald, Stuart, Wells, David and Mattheis, Ashley. (2024). "Using Artificial Intelligence and Machine Learning to Identify Terrorist Content Online". *Tech Against Terrorism Europe*, 15 January. <https://tate.techagainstterrorism.org/news/tcoaireport>
- Naseer, Muhammad Ansar, and Ghazala Shaheen. "Harnessing the Power of Artificial Intelligence: An In-Depth Review of its Effective Role in Countering Violent Extremism." *JURIHUM: Jurnal Inovasi dan Humaniora* 1, no. 4 (2023): 569-580.



- Schröter, Marie (2020). "Artificial Intelligence and Countering Violent Extremism: A Primer." *Global Network on Extremism and Technology*, 28 September. <https://gnet-research.org/2020/09/28/artificial-intelligence-and-countering-violent-extremism-a-primer/>
- Barnes, M.R., (2022). Online extremism, AI, and (human) content moderation. *Feminist Philosophy Quarterly*, 8(3/4).
- Fernandez, Miriam, and Harith Alani. (2021). "Artificial intelligence and online extremism: Challenges and opportunities." *Predictive policing and artificial intelligence*: 132-162.
- Heller, Brittan. (2020). "Combating terrorist-related content through AI and information sharing." *Algorithms*: 1-8.
- Tekiroglu, Serra Sinem, Helena Bonaldi, Margherita Fanton, and Marco Guerini. (2022). "Using pre-trained language models for producing counter narratives against hate speech: a comparative study." *arXiv preprint arXiv:2204.01440*
- Montasari, Reza. (2024). "Machine Learning and Deep Learning Techniques in Countering Cyberterrorism." In *Cyberspace, Cyberterrorism and the International Security in the Fourth Industrial Revolution: Threats, Assessment and Responses*, pp. 135-158. Cham: Springer International Publishing.
- Costello, Thomas H., Pennycook, Gordon and Rand, David G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science* 385, eadq1814 (2024). DOI:10.1126/science.adq1814. <https://www.science.org/doi/10.1126/science.adq1814>
- Gandhi, Milan. (2024). *Terrorism, Extremism, Disinformation and Artificial Intelligence: A Primer for Policy Practitioners*. Institute For Strategic Dialogue, 22 January. <https://www.isdglobal.org/isd-publications/terrorism-extremism-disinformation-and-artificial-intelligence-a-primer-for-policy-practitioners/>
- Huamani, Enrique Lee, Mantari Alicia Alva, and Avid Roman-Gonzalez. (2020). "Machine learning techniques to visualize and predict terrorist attacks worldwide using the global terrorism database." *International Journal of Advanced Computer Science and Applications* 11, no. 4.
- Khan, Fahad Ali, Gang Li, Anam Nawaz Khan, Qazi Waqas Khan, Myriam Hadjouni, and Hela Elmannai. (2023). "AI-Driven Counter-Terrorism: Enhancing Global Security Through Advanced Predictive Analytics." *IEEE Access* 11: 135864-135879.
- McKendrick, Kathleen. (2019). "Artificial intelligence prediction and counterterrorism." *London: The Royal Institute of International Affairs-Chatham House* 9.
- Ige, Tosin, Abosede Kolade, and Olukunle Kolade. (2022). "Enhancing border security and countering terrorism through computer vision: A field of artificial intelligence." In *Proceedings of the Computational Methods in Systems and Software*, pp. 656-666. Cham: Springer International Publishing.
- Stokes, Skyler. (2021). "The AI Revolution and Its Implications on Domestic Counterterrorism". *Middlebury Institute of International Studies at Monterey*, July 27. <https://www.middlebury.edu/institute/academics/centers-initiatives/ctec/ctec-publications/ai-revolution-and-its-implications-domestic#:~:text=While%20AI%20is%20certain%20to,come%20to%20harm%20innocent%20civilians%20>
- Jha, Anushka, Rajesh Bahuguna, Samta Kathuria, G. Sunil, Manish Gupta, and Vikrant Pachouri.(2023). "Role of AI in Combating Cyber Terrorism." In *2023 4th International Conference on Smart Electronics and Communication (ICOSEC)*, pp. 1156-1160. IEEE.
- Ceballos, Raquel Velasco. (2022). "AI in Military Affairs". *The European Land Force Commanders Organisation*, 8 November. <https://finabel.org/ai-in-military-affairs-its-role-in-the-decision-making-process-towards-a-counter-terrorism-operation/>
- UN Women. (2024). "Artificial intelligence and the Women, Peace, and Security Agenda in South-East Asia." *UN Women*. <https://asiapacific.unwomen.org/en/digital-library/publications/2024/05/artificial-intelligence-and-the-women-peace-and-security-agenda>
- Royer, Alexandrine. (2020). "The short anthropological guide to the study of ethical AI." *arXiv preprint arXiv:2010.03362*.
- High-level Advisory Body on Artificial Intelligence. (2024). "Governing AI for Humanity". *UN Advisory Body*, September. <https://www.un.org/en/ai-advisory-body>



- Burton, Joe. (2023). "Algorithmic extremism? The securitization of artificial intelligence (AI) and its impact on radicalism, polarization and political violence." *Technology in society* 75: 102262.
- Naseer & Shaheen, "Harnessing"; Burton, "Algorithmic"; High-level Advisory Body on Artificial Intelligence, "Governing AI"; Huamani, et al, "Machine learning", Macdonald, Wells & Mattheis, "Using"; Stokes, "AI Revolution".
- Conway, Maura. (2021). "Online extremism and terrorism research ethics: Researcher safety, informed consent, and the need for tailored guidelines." *Terrorism and political violence* 33, no. 2: 367-380.
- Reissman, Hayley. (2023). "What is Deepfake Porn and Why is Thriving in the Age of AI?" *Annenberg School for Communication, University of Pennsylvania*, 13 July. <https://www.asc.upenn.edu/news-events/news/what-deepfake-porn-and-why-it-thriving-age-ai>
- Hedayah (2024). *Expert Roundtable Discussions – Artificial Intelligence for Countering Extremism*.



ANNEX A. DETAILED METHODOLOGY

This Detailed Methodology outlines the methodology for this exploratory, qualitative study conducted over the course of 2024 by Hedayah's Research & Analysis Department with support from an external expert consultant.

Research Objectives & Questions

With a constantly evolving technology, and terrorist and extremist actors who are often adept early adopters of these technologies, there is an ongoing need to study new trends, anticipate new challenges, and further, to consider how these challenges can be addressed, and how AI may also be a positive tool to support CT and counter extremism work.

These developments highlighted a need to:

- Identify the threats that AI poses in terms of terrorist and extremist exploitation.
- Consider the opportunities that AI presents to work to counter and prevent terrorism, violent extremism, and extremism, as well as the potential challenges associated with utilizing AI in this manner.

In this context, Hedayah proposed a program of work including convening relevant experts in a neutral platform to discuss these emerging threats and potential opportunities, pinpointing needs for policy and programming and areas for further research, and conducting review of existing literature.

The research sought to answer the following high-level questions, in line with the Initiative's objectives:

- **Threats:** What are the key current threats that artificial intelligence, particularly generative AI, poses in terms of terrorist and extremist exploitation, and what new threats may be emerging?
- **Opportunities:** How could generative AI be used, or how is it already being used, to support efforts in countering and preventing extremism and violent extremism? What lessons can we learn from previous uses of (non-generative) AI for these purposes?
- **Challenges:** What are the potential ethical and practical challenges that may be associated with utilizing generative AI for prevention of extremism and violent extremism?
- **Needs:** What needs are emerging or expected in terms of utilizing generative AI for prevention or addressing associated challenges, and how might these needs be addressed?

The research also sought to consider the following cross-cutting issues:

- **Gender differentiated challenges:** How may factors of identity, primarily gender, but also age, background, etc. impact these threats, opportunities, challenges, and needs?
- **Non-Western perspectives:** How do current threats, opportunities, challenges and needs differ in different contexts, in particular in non-Western contexts?

Research Approach

Literature Review

To evaluate the key themes related to the exploitation of generative AI by terrorists and the opportunities it presents for counterterrorism, a targeted literature review was conducted. This review employed a combination of key terms alongside a manual selection process to ensure a balanced representation of



backgrounds, contexts, and approaches. Only articles that included "artificial intelligence" in their keywords — whether designated by the authors or identified by the author of this paper—were considered for inclusion. Consequently, some broader areas of research, such as the role of recommender systems in promoting terrorist content or the effects of algorithmic amplification on radicalization, were largely excluded from this study. This approach allowed for a focused analysis on the specific intersections of AI, terrorism, and counterterrorism, ensuring the literature reviewed was directly relevant to these themes. While Hedayah’s focus is on counter extremism rather than counterterrorism, less research spoke to extremism than terrorism, but does mention radicalization towards extremism and extremist content.

In total, **52 studies were analyzed**, making up the corpus for this research. Whereas some include academic research, others include government policy papers, thinktank analysis, and blogs. A full list can be found in the bibliography of this paper. All articles were written between 2018 and 2024. As the focus of this paper is assessing the four main pillars, the articles were structured according to theme. These included:

- **Threats:** Terrorist and extremist use of AI
- **Opportunities:** The use of AI for counterterrorism
- **Challenges:** Governance & ethics
- *Full review:* Used for both threats and opportunities

The table below shows the articles per theme:

Theme	Description	# of articles in scope
<i>The use of AI for counterterrorism</i>	Focusses on utilizing AI for counterterrorism such as drone usage and predicting attacks.	19
<i>Terrorist and extremist use of AI</i>	Focusses solely on the use of AI by terrorist and extremist actors, whether experimental or based on real events / evidence.	19
<i>AI Governance & Ethics</i>	Focusses on governance structures for ethical AI.	5
<i>Full review</i>	Includes a combination of the opportunities of AI for counterterrorism as well as the exploitation.	9

Roundtable Discussions

The Initiative began with a series of Roundtable Discussions of approximately 2 hours. These sessions were structured to allow discussions on a variety of different relevant topics, specifically relating to the research questions outlined above. These roundtables will not be grouped by stakeholder type or region, but will be scheduled to facilitate participation across different time zones (likely resulting in some natural geographic focus in individual sessions).

Potential partners, stakeholders and attendees targeted included:

- Counter extremism and counter terrorism tech actors
- Academic and research organizations
- Government: Policymakers who have developed frameworks in response to AI and its terrorist or extremist applications
- International organizations
- Tech sector: Tech companies who are already engaging or leading on the issues of risk related to AI



- Practitioners: Representatives of organizations already using AI in their CT/counter extremism programming

Ultimately, five **(5) roundtables were organized and conducted** with representatives from the practitioner, academic, tech, civil society, and governmental sectors. Experts were asked to specifically comment on the current threat landscape and exploitation of generative AI by terrorists and violent extremists, how AI can be used for counterterrorism or counter extremism efforts, what the challenges of doing so could look like, and what the current needs are to counter such exploitation.

Participating Organizations	
Global Internet Forum to Counter Terrorism (GIFCT)	Swansea University / VOX-Pol
University of Silesia	Indonesia Knowledge Hub (iK-Hub) on Countering Terrorism and Violent Extremism
Paperlab	University of Wisconsin-Madison
Moonshot	Southeast Asia Regional Centre for Counter Terrorism (SEARCCT)
CIVIPOL/Hedayah	Deakin University
University of Swansea	Extremism and Gaming Research Network (EGRN)
Counter-Terrorism Committee Executive Directorate (UN CTED)	Modulate.ai
Anglia Ruskin University	Centinel
Institute for Strategic Dialogue	


Please note that roundtables were held under the Chatham House Rule, and as such, no individuals are quoted in this Brief. We thank all our participants for their insights and for shaping this paper.







Hedayah
Countering Extremism
& Violent Extremism

www.hedayah.com

 [hedayah_CVE](https://twitter.com/hedayah_CVE)

 facebook.com/hedayah

 linkedin.com/company/hedayah

 [hedayah_cve](https://instagram.com/hedayah_cve)